Pricing flexibility of shiftable demand in electricity markets

Lucien Werner
lwerner@caltech.edu
California Institute of Technology
Pasadena, CA, USA

Adam Wierman
adamw@caltech.edu
California Institute of Technology
Pasadena, CA, USA

Steven H. Low
slow@caltech.edu
California Institute of Technology
Pasadena, CA, USA

ABSTRACT

Enabling participation of demand-side flexibility in electricity markets is key to improving power system resilience and increasing the penetration of renewable generation. In this work we are motivated by the curtailment of near-zero-marginal-cost renewable resources during periods of oversupply, a particularly important cause of inefficient generation dispatch. Focusing on shiftable load in a multi-interval economic dispatch setting, we show that incompatible incentives arise for loads in the standard market formulation. While the system's overall efficiency increases from dispatching flexible demand, the overall welfare of loads can decrease as a result of higher spot prices. We propose a market design to address this incentive issue. Specifically, by imposing a small number of additional constraints on the economic dispatch problem, we obtain a mechanism that guarantees individual rationality for all market participants while simultaneously obtaining a more efficient dispatch. Our formulation leads to a natural definition of a uniform, time-varying flexibility price that is paid to loads to incentivize flexible bidding. We provide theoretical guarantees and empirically validate our model with simulations on real-world generation data from California Independent System Operator (CAISO).

CCS CONCEPTS

• **Hardware** → **Renewable energy**; • **Mathematics of computing** → *Mathematical analysis.*

KEYWORDS

electricity markets, demand response, mechanism design, incentives, flexibility, shiftable demand

ACM Reference Format:
Lucien Werner, Adam Wierman, and Steven H. Low. 2021. Pricing flexibility of shiftable demand in electricity markets. In *The Twelfth ACM International Conference on Future Energy Systems (e-Energy '21), June 28-July 2, 2021, Virtual Event, Italy*. ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3447555.3464847

1 INTRODUCTION

The traditional paradigm of generation following load is being transformed as variable, non-dispatchable resources like solar and wind are an ever increasing share of the generation mix. This creates situations, typically during midday, where there is an excess near-zero marginal cost renewable generation which must be curtailed in order to maintain supply-demand balance. While this scenario might have seemed far-fetched even a few years ago, it is already occurring in major markets. On October 11, 2020, renewables met more that 100% of the total demand in Southern Australia for several hours [3]. In the CAISO (California Independent System Operator) market, solar regularly provides more that 60% of total generation during the afternoon and reached an all-time peak of 80% in May 2019 [12]. Due to a generation queue dominated by renewables and a 100% zero-emission target for 2045, over-generation from renewables will become increasingly common in the California and other large markets [1].

Storage and demand response are two approaches for shifting surplus renewable generation from peak midday hours to periods of higher demand. In an influential 2017 report, researchers at Lawrence Berkeley National Lab analyzed opportunities for demand response and proposed the "Shape-Shift-Shed-Shimmy" taxonomy of flexible loads [7]. They argue that each type of load flexibility is applicable for a particular timescale and use case. Shift flexibility, where the total energy consumed over the time horizon (e.g., 24 hours) remains constant but can be shifted between time intervals, is identified as the form of demand response best suited to accommodate renewable over-generation. Sources of shiftable load include electric vehicle charging, commercial and residential HVAC, and non-time-sensitive industrial processes.

Mechanisms and incentives for offering demand response have been extensively studied but most often they focus on direct compensation for load shedding or peak shaving. Demand response programs implemented by ISOs (Independent System Operators) and utilities tend to be tailored to that same goal. Despite the calls for attention, mechanisms for Shift flexibility in particular remain relatively understudied [7, 37]. The operational benefits of dispatching shiftable loads are clear to market operators, but as existing markets do not invite significant demand-side participation, from the consumer's point of view the advantages are less clear. This motivates the core questions of this work: *Is a flexible load better off offering its shiftable demand to the market operator than not? And, if not, can we redesign the market to encourage loads to offer shiftable demand to the marketplace?*

Contributions. The answers are No and Yes. We prove that there is incentive misalignment in traditional market designs where flexible loads may prefer not to expose their flexibility to the marketplace. To address this, we introduce a new mechanism where loads have incentives to offer flexibility and generator incentives remain aligned with the social welfare objective. More specifically, this paper makes the following contributions.

e-Energy '21, June 28-July 2, 2021, Virtual Event, Italy

ACM ISBN 978-1-4503-8333-2/21/06.
https://doi.org/10.1145/3447555.3464847

First, we establish a market and utility model for analyzing shiftable demand. Ours is a variant of the multi-interval market, extensively studied with ramping inequality constraints [26, 29, 31, 48] where equality constraints are added to couple the demand consumption in all periods. Our framework for load utility is derived from the load utility model implied by the standard economic dispatch formulation.

Second, we identify a fundamental incentive incompatibility for loads offering flexible dispatch while being compensated with the standard electricity spot price. We show in Theorem 3.1 that even in very simple scenarios, loads are worse off under flexible dispatch, even as generators capture more profit and the efficiency of the dispatch solution improves. This counter-intuitive situation arises from the interplay between the time-coupling demand constraint and the power balance constraint that holds in each time interval.

Third, we propose a new multi-interval economic dispatch market that corrects the demand-side incentive incompatibility. Our mechanism preserves core features of the existing structure while making some novel changes: we add inequality constraints to constrain the demand allocation and clear the generation and demand sides of the market separately in a two-step procedure that ensures supply-demand balance and revenue adequacy. Loads that offer flexibility are compensated for deviating from their nominal baseline with a flexibility price, defined in Section 4, while inflexible loads continue to pay the baseline spot price for energy. Our main result, Theorem 4.3, proves that loads have incentives to offer flexibility under the new market design without disturbing dispatch-following incentives for generators.

Finally, in Section 5 we present a case study using generation data from CAISO. The case study highlights the importance of ensuring that shiftable loads have incentives to bid their flexibility into the marketplace. Our results show that curtailment of renewable generation can be eliminated, leading to a 15% reduction in the net generation costs.

Related work. This paper builds on and contributes to three areas of the literature on electricity markets: (1) mechanisms for demand response, (2) multi-interval dispatch, and (3) incentive alignment in mechanism design.

Demand response has been extensively explored in both the academic literature and in practice. In both contexts, interest in demand response has mainly centered around rate-based demand reduction [19, 22, 32, 45] and incentive-based programs [13]. In the former category are time-of-use pricing [18], critical-peak pricing [27], and real-time pricing [6], all schemes which use a given price schedule to incentivize loads to consume energy during lower-cost periods. In the latter group are programs like direct load control [21] and emergency demand reduction [4] in which loads are given lump-sum or per-event payments by the system operator in exchange for the curtailment. Such programs are popular in practice since they lower demand and spot prices during peak load hours.

A drawback of many of these existing designs is that they tend to emphasize a particular variety of demand response—load shedding—and do not explicitly offer incentives for other types of flexible load. The demand response taxonomy in [7] identifies four major types of demand response, each requiring their own dispatch and incentive structures. A general mathematical formulation for optimal dispatch of flexible load is notably given in [37] but the formulation therein assumes knowledge of demand-side value functions. In practice these are very difficult to determine, partly for practical reasons (there are seldom opportunities for loads to reveal their price elasticity) and partly due to historical reasons (electricity has always been treated as an "on-demand" commodity) [33]. We are not aware of any works that formally analyzes incentives for offering shiftable demand—identified as the most significant potential source of demand-side flexibility in [7]—while also retaining the established economic dispatch market structure.

Another important theme in the demand response literature is strategic behavior by loads when reporting their baseline energy consumption. Because demand response almost always defined as a reduction from a baseline, there can be incentives for loads to inflate their baselines to give the appearance of a larger load reduction in real-time. There are a number of works that analyzed incentives for misreporting and proposed mechanisms to discourage it [14, 16, 39, 40, 46]. While we retain the concept of a baseline in this work for convenience, our model is compatible with schemes to limit the incentives to misreport it, e.g., [41].

Multi-interval markets are of growing interest as a way to guarantee reliable electricity dispatch in the face of uncertain generation. Several substantial works have explored multi-interval market design including [26, 29, 31, 48]. The intertemporal constraints in all of these are limited to ramping limits, which only couple adjacent time periods. In contrast, along the lines of the model proposed in [37], our work considers equality constraints on demand consumption that couple all time periods together. This type of inter-temporal constraint introduces a particular incentive misalignment—a the focus of this paper.

More broadly, our work connects to the topic of mechanism design. Analysis of incentive and participation constraints in market mechanisms was pioneered by Hurwicz, Groves, and Ledyard, among others [25, 34]. The study of incentives in electricity markets has a rich history beginning with the seminal work of Schweppe [44] and has strongly influenced subsequent research on congestion pricing [15, 28] and non-convex pricing [8, 24, 30, 36]. In addition there has been research on market manipulation by generators, e.g., through market power and/or strategic curtailment of renewable generation. Some notable recent results in this direction include [10, 35, 42, 43]. While this body of work establishes a framework for analyzing electricity market incentives, it does so almost exclusively for the generation side of the market [38]. Efficiently dispatching demand-side resources to meet system needs requires similar evaluation of incentive structures.

2 MARKET MODEL

We study an economic dispatch market for energy that the market operator (e.g., ISO/RTO) uses to calculate dispatch quantities and settlement prices. Our model is distinctive from the standard short-term setting in several important ways. First, we consider a multi-interval market with intertemporal equality constraints, which are necessary to model shiftable demand. This contrasts with an existing body of work on multi-interval markets with intertemporal inequality constraints. Second, we explicitly model and dispatch the demand side of the market. Typically demand is taken to be

fixed with only generation being variable. Third, we evaluate the welfare of *both* generators and loads in our analysis of incentives. As loads are the participants providing demand response flexibility, explicitly incorporating them into the social welfare formulation is crucial for quantifying the impacts of flexibility.

2.1 Market participants

The market has N generators, indexed by i, and operates over discrete time horizon of length T, indexed by t. The energy produced by generator i in interval t is denoted by $p_{i,t} \in \mathbb{R}$. We denote generators' production over the time horizon with the generation matrix $\mathbf{P} \in \mathbb{R}^{N \times T}$. It is sometimes convenient to refer to individual row/columns of this matrix. The t-th column, the market production vector for time t, is $\mathbf{p}_t = [p_{1,t}, \ldots, p_{N,t}]^\top \in \mathbb{R}^N$. Analogously, the i-th row, generator i's production across the entire time horizon, is denoted $\mathbf{p}_i = [p_{i,1}, \ldots, p_{i,T}]^\top \in \mathbb{R}^T$. Generator cost functions $c_{i,t}(p_{i,t}) : \mathbb{R} \to \mathbb{R}_+$ are assumed to be convex, monotonically increasing, sub-differentiable, and zero-crossing. This last property requires that $0 \in \mathbf{dom}(c_{i,t})$ and $c_{i,t}(0) = 0$. For convenience, we refer to the total cost function for each generator as

$$c_i(\mathbf{p}_i) = \sum_t c_{i,t}(p_{i,t}).$$

The market includes M demand participants, which we refer to as loads, indexed by j. Each load consumes a fixed amount of energy E_j over the T periods. We use $d_{j,t} \in \mathbb{R}$ to denote the the energy consumed by load j in interval t. Like with generators, we stack the $d_{j,t}$ into a demand matrix $\mathbf{D} \in \mathbb{R}^{M \times T}$. We refer to the t-th column with $\mathbf{d}_t = [d_{1,t}, \ldots, d_{M,t}]^\top \in \mathbb{R}^M$ and the j-th row with $\mathbf{d}_j = [d_{j,1}, \ldots, d_{j,T}]^\top \in \mathbb{R}^T$. Loads do not have preference functions that vary with consumption in each time interval. However, they do report a preferred baseline in each interval $d^0_{j,t} \in \mathbb{R}$. E_j is defined in terms of the cumulative baseline consumption of load j:

$$E_j = \sum_t d^0_{j,t}.$$

The use of a baseline is a common assumption in demand response (see e.g., [45]) that we retain here in order to provide a natural definition of flexibility Δ_j as the amount that the actual dispatch $\mathbf{d}_j$ deviates from load j's preferred baseline $\mathbf{d}^0_j$: $\Delta_j := \mathbf{d}_j - \mathbf{d}^0_j$.

2.2 Market mechanism

The structure of bids, the market clearing procedure, and the settlement structure are laid out in the following steps:

1. All participants (loads, generators) submit their bids. For loads, this takes the form of a triple
$$(E_j, \underline{\mathbf{d}}_j, \overline{\mathbf{d}}_j) \in \mathbb{R}_+ \times \mathbb{R}^T_+ \times \mathbb{R}^T_+$$
that consists of their energy requirement and lower/upper bounds on consumption in each time period.[1] For generators, the bid takes the form of a pair
$$(c_i, \overline{\mathbf{p}}_i) \in C \times \mathbb{R}^T_+$$
C is the set of all functions $c : \mathbb{R}^T_+ \to \mathbb{R}$ that are convex, monotonically increasing, and contain the origin. Generators only submit their upper bounds on production; to avoid non-convexities arising from unit commitment, generation lower bounds are assumed to be 0.
2. The market operator collects bids and solves a market clearing optimization problem, defined in (1a) - (1e). Its solution provides an allocation of energy to each participant (the dispatch) and a unit price for energy in each time period.
3. Generators are obligated to produce the dispatch quantities and are paid the unit price for whatever they produce. Loads must consume the dispatch quantities and must pay the unit price for whatever they consume. If any participant deviates from the dispatch schedule, the market operator has the ability to administratively penalize the violator, e.g., via large monetary penalties or exclusion from the market.[2]

The centerpiece of the market structure is the *market clearing optimization problem* in Step 2. We study a version of the economic dispatch problem used by ISOs, made distinctive in our case by the multi-interval setting and the inclusion of intertemporal equality constraints. For the sake of focusing our analysis on the impacts of these unique features, we do not consider unit commitment, start-up/no-load costs, and line congestion. We also ignore ramping constraints (i.e., intertemporal inequality constraints) for both loads and generators. As previously mentioned, these have been studied extensively on the generation side of market in [26, 29, 31, 48] among others. Finally, we consider a "single-shot" market-clearing procedure where dispatch quantities and prices are determined at the beginning of the dispatch horizon and adhered to through the remainder of it.[3] Relaxing these simplifying assumptions is discussed as future work in the Conclusion but we note here that incentive misalignment for loads arises even in the most straightforward setting of the problem.

The market clearing optimization problem is as follows:

$$\min_{\mathbf{p}_j, \mathbf{d}_j \ \forall i,j} \quad \sum_i c_i(\mathbf{p}_i) \tag{1a}$$

subject to

$$\lambda_t \ \perp \ \mathbf{1}^\top \mathbf{d}_t - \mathbf{1}^\top \mathbf{p}_t = 0 \qquad \forall t \tag{1b}$$

$$\rho_j \ \perp \ \mathbf{1}^\top \mathbf{d}^j = E_j \qquad \forall j \tag{1c}$$

$$\mu^-_i, \mu^+_i \ \perp \ \mathbf{0} \le \mathbf{p}_i \le \overline{\mathbf{p}}_i \qquad \forall i \tag{1d}$$

$$\eta^-_j, \eta^+_j \ \perp \ \underline{\mathbf{d}}_j \le \mathbf{d}_j \le \overline{\mathbf{d}}_j \qquad \forall j \tag{1e}$$

In the above, (1a) is the total generation cost; (1b) are the power balance constraints in each interval; (1c) enforces that each load's energy requirement E_j is met over the time horizon (these are the intertemporal equality constraints); and (1d) - (1e) ensure that the dispatch satisfies participants' minimum and maximum production/consumption limits.

Given an optimal solution to (1), the time-varying non-negative energy price π_t is defined for all t as

$$\pi_t := \lambda^*_t \tag{2}$$

where λ^*_t is the optimal dual variable for (1b).

[1] As $\mathbf{d}_j$ is assumed to be non-negative, all components of the bid are non-negative.

[2] This requirement reflects the auction design of most North American ISOs, see e.g., Section 2.1 in [17].

[3] The only assumption needed to support this is that the $\mathbf{d}^0_j$ are known at $t = 1$ and do not adjust over the course of the time horizon.

By offering flexibility in the form of a box constraint on demand as in (1e), the efficiency of the dispatch is improved. This is expressed in the following theorem.

Theorem 2.1. *Let* OPT^0 *be the optimal value of problem (1a) - (1e) when* $\underline{d}_{j,t} = \overline{d}_{j,t} = d^0_{j,t}$ *for all* j *and* t*, assuming it exists. Let* OPT *be the optimal value of the problem where* $\underline{d}_{j,t} < d^0_{j,t} < \overline{d}_{j,t}$ *for at least one* j *or* t*. Then* $\text{OPT} \leq \text{OPT}^0$.

Theorem 2.1 states that dispatching demand-side flexibility offers benefits to the market allocation in the form of lower cost (greater efficiency). A proof of this result follows immediately from the fact that relaxing constraint (1e) results in a large feasible set, which therefore gives a lower bound on the optimal value in the case where the constraint is tight. The existence of OPT is guaranteed by the existence of OPT^0.

Notice that the formulation of economic dispatch in (1) reduces to the standard setting (T independent sequential economic dispatch problems) when $\underline{\mathbf{d}}_j = \mathbf{d}_j = \overline{\mathbf{d}}_j$. In this situation, constraints (1c) and (1e) are redundant and $\mathbf{d}_t$ in (1b) can be replaced by $\mathbf{d}^0_t$.

2.3 Utility models for generators and loads

An important goal of this paper is to evaluate whether the market allocation, given by the optimal primal solution of (1), and the market-clearing price, given by the optimal dual variable of (1b), are aligned with the individual incentives. To study this question we need to introduce definitions of utility and the individual utility maximization problem for both loads and generators.

Let $\pi \in \mathbb{R}^T_+$ be the vector of market-clearing energy prices for the time horizon. We assume all agents are price takers and define the following utility models for generators and loads respectively.

Definition 2.2. Let $\mathbf{p}_i \in \mathbb{R}^T$ *be generator* i*'s production vector and* $\mathcal{P}_i$ *be its private constraint set*

$$\mathcal{P}_i = \left\{ \mathbf{p}_i \in \mathbb{R}^T \;\middle|\; \mathbf{0} \leq \mathbf{p}_i \leq \overline{\mathbf{p}}_i \right\} \subseteq \mathbb{R}^T.$$

Generator i*'s utility is defined as*

$$u_i\left(\mathbf{p}_i; \pi\right) := \pi^\top \mathbf{p}_i - c_i(\mathbf{p}_i). \tag{3}$$

We assume a generator acts rationally when facing the given price schedule π and therefore seeks to maximize its utility with

$$\begin{aligned} \arg\max_{\mathbf{p}_i} \quad & u_i\left(\mathbf{p}_i; \pi\right) \\ \text{s.t.} \quad & \mathbf{p}_i \in \mathcal{P}_i. \end{aligned} \tag{4}$$

In contrast to generators, loads do not have a cost function and are only constrained by a required amount of energy to be delivered over the time horizon, E_j. Instead we assume that there is a constant utility value $U_j \in \mathbb{R}_+$ that represents the value a load receives from having E_j satisfied. We assume that the load is indifferent to how energy is allocated across the intervals, as long as E_j is delivered and upper/lower consumption limits are respected.

Definition 2.3. Let $\mathbf{d}_j \in \mathbb{R}^T$ *be load* j*'s consumption vector and* $\mathcal{D}_j$ *be its private constraint set,*

$$\mathcal{D}_j = \left\{ \mathbf{d}_j \in \mathbb{R}^T \;\middle|\; \underline{\mathbf{d}}_j \leq \mathbf{d}_j \leq \overline{\mathbf{d}}_j, \quad \mathbf{1}^\top \mathbf{d}_j = E_j \right\} \subseteq \mathbb{R}^T.$$

Load j*'s utility is defined as*

$$u_j\left(\mathbf{d}_j; \pi\right) = U_j - \pi^\top \mathbf{d}_j. \tag{5}$$

We again assume each load acts rationally when facing the given price schedule π and therefore seeks to maximize its utility with

$$\begin{aligned} \arg\max_{\mathbf{d}_j} \quad & u_j(\mathbf{d}_j; \pi) \\ \text{s.t.} \quad & \mathbf{d}_j \in \mathcal{D}_j. \end{aligned} \tag{6}$$

A feature of this presentation of utility that deserves comment and justification is our representation of the positive "value" component of load utility with a constant U_j. This choice is made to align with the classical auction-based economic dispatch model in Section 2.2. Specifically, if we use utility functions (3) and (5) to construct the market's social welfare maximization problem subject to a shared market clearing constraint and private feasibility constraints, we get exactly the auction-based economic dispatch model (i.e., cost minimization) of the market described by (1). To see this, recall that the market's social welfare maximization problem is

$$\begin{aligned} \max_{\mathbf{p}_i, \mathbf{d}_j \;\forall i,j} \quad & \sum_i u_i(\mathbf{p}; \pi) + \sum_j u_j(\mathbf{d}_j; \pi) & \\ \text{s.t.} \quad & \mathbf{1}^\top \mathbf{p}_t = \mathbf{1}^\top \mathbf{d}_t & \forall t \\ & \mathbf{p}_i \in \mathcal{P}_i & \forall i \\ & \mathbf{d}_j \in \mathcal{D}_j & \forall j \end{aligned} \tag{7}$$

At an optimal point $\mathbf{p}^*_i, \mathbf{d}^*_j \;\forall i, j$ of (7),

$$\begin{aligned} & \mathbf{1}^\top \mathbf{p}^*_t - \mathbf{1}^\top \mathbf{d}^*_t = 0 \quad \forall t \\ \Rightarrow \; & \sum_i \mathbf{p}_i - \sum_j \mathbf{d}_j = \mathbf{0} \end{aligned}$$

Plugging in (3) and (5) into the objective function of (7), we get

$$\begin{aligned} \sum_i u_i(\mathbf{p}; \pi) + \sum_j u_j(\mathbf{d}_j; \pi) &= \sum_i \pi^\top \mathbf{p}_i - c_i(\mathbf{p}_i) + \sum_j U_j - \pi^\top \mathbf{d}_j \\ &= \sum_j U_j - \sum_i c_i(\mathbf{p}_i) + \pi^\top \Big(\sum_i \mathbf{p}_i - \sum_j \mathbf{d}_j \Big) \\ &= \sum_j U_j - \sum_i c_i(\mathbf{p}_i) \end{aligned}$$

It is clear that the objective function of (7) differs from (1a) by only a constant factor and the constraint sets of the two problems are identical. Therefore, they have the same optimal solution (although the optimal value differs by a factor of $\sum_j U_j$). While the choice of U_j does not impact the optimal solution, intuitively, it should be a positive number, large enough so that $U_j - \pi^\top \mathbf{d}_j > 0$ for most realizations of π and $\mathbf{d}_j$. However this condition is not necessary for our analysis of prices and dispatch quantities.

3 PARTICIPATION INCENTIVES

Our first set of results focuses on understanding the consequences of dispatching flexibility via the classical market formulation described in the previous section. We show in this section that, though dispatch-following incentives for generators remain intact, participation incentives for loads are misaligned and offering flexibility (i.e., $\underline{\mathbf{d}}_j < \overline{\mathbf{d}}_j$) to the market operator is not necessarily rational.

3.1 Participation incentives for loads

Participation constraints affect a rational agent's behavior. In particular, given a choice to enter into a market/mechanism or not, it is expected that a rational agent only does so if their utility is higher under participation than their best alternative. To put this precisely for the case of loads in our model, let $\mathbf{d}_j^0$ be the allocation a load receives outside of the flexibility mechanism (i.e., the load simply consumes its reported baseline). Let $\mathbf{d}'_j$ be the allocation a load receives by participating in the mechanism. j's participation constraint is satisfied if and only if $u_j(\mathbf{d}'_j) \geq u_j(\mathbf{d}_j^0)$.

Once a participant submits its bid to the market operator, it is obliged to obey the dispatch instruction that comes in return when the market is cleared. We show in the following theorem that, depending on the market outcome, loads can end up worse off by offering flexibility under the energy price in (2), despite the increase in efficiency that flexibility offers to the market as a whole (established in Theorem 2.1).

Theorem 3.1. *Assume the baseline solution $\mathbf{d}_j^0\ \forall j$ is feasible for (1a) - (1e). Under the market dispatch (1a) - (1e) and energy price π_t given in (2), participation constraints for loads are are not guaranteed. That is, there exist choices of parameters $c_i(\cdot), \mathbf{d}_j^0, \overline{\mathbf{p}}_i, \underline{\mathbf{d}}_j, \overline{\mathbf{d}}_j$ with $\underline{d}_{j,t} < \overline{d}_{j,t}$ for some j, t such that $u_j(\mathbf{d}'_j) < u_j(\mathbf{d}_j^0)$ for some j.*

Proof. Our proof takes the form of a counterexample. Consider a market environment with 2 time periods: $t = 1, 2$. There is a single load with demand given by $\mathbf{d} = [d_1, d_2]^\top$ and a single generator with generation given by $\mathbf{p} = [p_1, p_2]^\top$. The unit generation costs are $\mathbf{c} = [1, 2]^\top$ and the baseline demand is $\mathbf{d}^0 = [2, 2]^\top$. Thus $E = 4$. Generation is constrained by $\underline{\mathbf{p}} = [0, 0]^\top, \overline{\mathbf{p}} = [3, 3]^\top$, and demand by $\underline{\mathbf{d}} = \mathbf{d}^0(1-\alpha) \leq \mathbf{d}^0(1+\alpha) = \overline{\mathbf{d}}$, where $\alpha \in [0, 1]$. We parameterize the demand lower/upper bounds with the constant α to allow us to vary the offered flexibility between 0 ($\alpha = 0$) and its maximum ($\alpha = 1$).[4]

Market dispatch model (1a) - (1e) with these parameters gives the following optimization problem:

$$\begin{aligned} \min_{p,d} \quad & c^\top \mathbf{p} \\ \text{s.t.} \quad & \lambda \perp \mathbf{p} = \mathbf{d} \\ & \mathbf{1}^\top \mathbf{d} = 4 \\ & \mathbf{0} \leq \mathbf{p} \leq 3 \cdot \mathbf{1} \\ & 2(1-\alpha) \cdot \mathbf{1} \leq \mathbf{d} \leq 2(1+\alpha) \cdot \mathbf{1}. \end{aligned} \tag{8}$$

By (2) the energy price vector is $\pi = \lambda^*$. We assume the value constant for the load is 0 and take the optimal solution of (8) to be $(\mathbf{p}', \mathbf{d}', \lambda')$, we have the following form of load utility:

$$u(\mathbf{d}') = -\lambda'^\top \mathbf{d}'.$$

We solve (8) for $\alpha \in [0, 1]$ and compute demand utility $u(\mathbf{d}'; \alpha)$. Since α parameterizes the "amount" of flexibility demand offers, increasing values of α correspond to greater demand flexibility (looser bounds on min/max consumption in each interval). The results are shown in Figure 1.

[4] $\alpha = 1$ is the maximum because demand cannot be negative. The upper bound does not have the same restriction as the lower one but we stick to a single parameter here for simplicity.

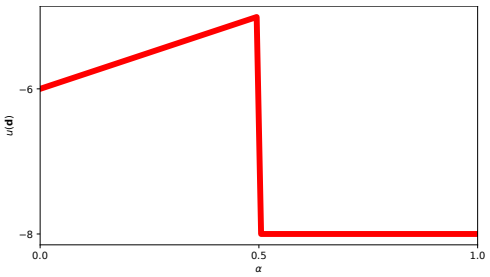


Figure 1: Demand utility vs. $\alpha \in [0, 1]$

Maximum consumer utility of -5 is reached as $\alpha \uparrow 0.5$ and $u(\mathbf{d}^0) \geq u(\mathbf{d})$ only for $\alpha < 0.5$. That is, the demand is worse off by offering for $\alpha \geq 0.5$ than none at all. In fact, if we had chosen the parameters differently (e.g., d^0 arbitrarily close to $\overline{p}$ in one interval), the demand participation constraint is violated for all $\alpha > 0$. □

Remark: *We retain the standard price-taking assumption in the above proof. With a single generator, this may be practically unrealistic. At the expense of greater complexity additional generators can be considered without changing the qualitative behavior we highlight. The purpose of the proof is to demonstrate that incentive violations arise even in the simplest of market settings.*

Analyzing which generation constraints in (1d) bind as α varies in the counterexample above gives insight into how misaligned incentives for loads come about.

Analogously to how a marginal generator $\hat{i}$ can be defined in the single-period economic dispatch, we define a marginal pair: generator *and* interval $(\hat{i}, \hat{t})$. If $\overline{p}_{\hat{i},\hat{t}} - \mathbf{1}^\top \mathbf{d}_{\hat{t}}^0 \geq 0$, then flexible demand can shift to $\hat{t}$ from costlier intervals to take advantage of this excess supply without changing the price $\lambda_{\hat{t}}$. However, once the upper bound is exceeded for the marginal pair (i.e., $\mathbf{1}^\top \mathbf{d}_{\hat{t}}^0 > \overline{p}_{\hat{i},\hat{t}}$), $\lambda_{\hat{t}}^*$ jumps up to the marginal cost of the next cheapest marginal pair. This surprising behavior occurs because time periods are coupled together through constraint (1c) Adding constraints on $\mathbf{d}_j$ that prevent this jump motivates the mechanism proposed in Section 4.

3.2 Participation incentives for generators

The previous section addresses the incentive misalignment for loads under the standard market structure. One may worry that a similar misalignment happens for generators. In this section we show that there is no such issue on the generator side of the market, i.e., that utility-maximizing decisions of the generators exactly match the dispatch decision by the market operator.

Specifically, the following theorem states that the optimal solution of the market dispatch problem (1a) - (1e) provides dispatch following incentives to generators, provided we treat generators as pricetakers. Along the way, we also show that generators do not have negative profit (i.e., participation constraints are satisfied). Throughout, we make the standard assumption that (1a) - (1e) has a feasible point.

Theorem 3.2. *Let $\mathbf{p}_i^*, \mathbf{d}_j^*\ \forall i, j$ be the optimal primal solutions of (1a) - (1e). The energy prices are $\pi := \lambda^*$ where λ^* is the vector of*

optimal dual variables for constraint (1b). Then

$$\mathbf{p}_i^* = \arg\max_{\mathbf{p}_i} \quad u_i(\mathbf{p}_i; \pi) \quad \text{s.t.} \quad \mathbf{p}_i \in \mathcal{P}_i$$

Further, $u_i(\mathbf{p}_i^*) \geq 0$.

Proof. See Appendix A. □

Note that this theorem extends a well-known result for single-period economic dispatch to our multi-interval setting with equality constraints.

4 INCENTIVIZING FLEXIBILITY

Section 3 highlights that loads have an incentive not to reveal their flexibility under the standard market design where only the energy price is used for settlement. This is problematic since exploiting the flexibility of loads is essential for system reliability, avoiding curtailment of renewable energy, and improving the economic efficiency of the dispatch. This section presents the main contribution of the paper: a new market design that ensures both loads and generation have incentives that are aligned with the market operator's and, specifically, provides incentives for loads to reveal their flexibility to the market. First we introduce the market design and prove its incentive properties and following, in Section 5, we illustrate the market design using a case study.

4.1 A market design for flexibility

Our proposed design adopts a similar structure to the standard market while introducing three important components: (1) a small number of additional constraints on the demand allocation, (2) a time-varying price κ_t for flexibility, and (3) a two-stage market clearing scheme for the demand side of the market.

Before presenting the mechanism we must introduce some notation. First, let the constant

$$c_{\min} := \min_{i,t} \frac{\partial c_{i,t}}{\partial p_{i,t}}(p_{i,t}^0)$$

be the smallest marginal cost (over all i and t) under the baseline allocation. Second, define $\mathcal{T} \subseteq \{1, \ldots, T\}$ to be the subset of intervals for which it is true that $\frac{\partial c_{i,t}}{\partial p_{i,t}}(p_{i,t}^0) = c_{\min}$ for at least one $i \in \{1, \ldots, N\}$. $\mathcal{T}^c$ is the set of all intervals that do not meet this condition. Together, $\mathcal{T} \dot\cup \mathcal{T}^c = \{1, \ldots, T\}$.[5] In what follows, we assume that neither $\mathcal{T}$ and $\mathcal{T}^c$ is empty. Third, for each $t \in \mathcal{T}$, define a generator index set

$$\mathcal{I}_t := \{i \mid \frac{\partial c_{i,t}}{\partial p_{i,t}}(p_{i,t}^0) = c_{\min}\} \subseteq \{1, \ldots, N\}.$$

Fourth, define

$$P_t^{\text{cap}} := \sum_{i \in \mathcal{I}_t} p_{i,t}^{\text{cap}}$$

where

$$p_{i,t}^{\text{cap}} = \arg\max_{p \in \mathbb{R}} p \quad \text{s.t.} \quad \frac{\partial c_{i,t}}{\partial p_{i,t}}(p) = c_{\min}, \quad p \leq \overline{p}_{i,t}$$

This (regrettably heavy) notation makes precise the amount of available *excess* capacity at the lowest price $c_{\min}$ in the baseline dispatch. Observe that when the c_i are linear, $p_{i,t}^{\text{cap}} = \overline{p}_{i,t}$ for $i \in \mathcal{I}_t$.

With this notation in hand, we summarize the structure of the market mechanism. Additional discussion of each step is provided following the exposition of the procedure.

1. Generators submit bids $(c_i, \overline{\mathbf{p}}_i)$ and loads submit bids $(\mathbf{d}_j^0, \underline{\mathbf{d}}_j, \overline{\mathbf{d}}_j)$ to the market operator.
2. Market operator collects bids, forms the market-clearing optimization problem (1a) - (1e) with the additional constraint $\mathbf{d}_j = \mathbf{d}_j^0$, and produces a baseline solution $(\mathbf{p}_i^0, \mathbf{d}_j^0, \pi^0)$ $\forall i, j$.
3. Market operator re-solves (1a) - (1e) with the addition of three new constraints:

$$\mathbf{1}^\top \mathbf{d}_t \leq P_t^{\text{cap}} \quad \forall t \in \mathcal{T} \tag{9a}$$

$$d_{j,t} \geq d_{j,t}^0 \quad \forall j, \ \forall t \in \mathcal{T} \tag{9b}$$

$$d_{j,t} \leq d_{j,t}^0 \quad \forall j, \ \forall t \in \mathcal{T}^c \tag{9c}$$

   An interim solution and prices are computed: $(\tilde{\mathbf{p}}_i, \tilde{\mathbf{d}}_j, \tilde{\pi})$ $\forall i, j$.
4. The market operator defines a flexibility price

$$\kappa = \kappa(\mathbf{d}_j^0, \tilde{\mathbf{d}}_j, \pi^0, \tilde{\pi}) \in \mathbb{R}^T \tag{10}$$

   as a function of optimal solutions of the two market clearing problems. (We discuss the precise form of κ in Section 4.3.)
5. The market operator solves a demand dispatch problem (11a) - (11f), producing a final allocation for demand: $\mathbf{d}_j^*$ $\forall j$.
6. Generators are dispatched to produce energy $\tilde{\mathbf{p}}_i$ at price $\tilde{\pi}$. Loads are dispatched to consume energy $\mathbf{d}_j^*$ at price π^0 and contribute flexibility Δ_j^* compensated with price κ.

We now walk through the steps in more detail, beginning with Step 2. Step 2 establishes a baseline allocation that is used later in the procedure to ensure that participation constraints are satisfied.

In Step 3, additional inequalities (9a) - (9c) constrain the demand dispatch to a desirable region. (9a) enforces that the total demand does not exceed the total maximum capacity of the cheapest generator(s) in the interval—provided that there is spare capacity under the baseline solution. (9b) ensures that demand can only *increase* if there is excess cheapest generation in a period. (9c) guarantees that demand can only *decrease* during intervals where all of the cheapest generation is already dispatched. These additional linear inequalities only add $|\mathcal{T}| + T$ constraints to the market dispatch problem, which already has $(1 + M + N)T + M$ constraints. Due to the assumption that $\mathcal{T}$ and $\mathcal{T}^c$ are non-empty, a solution to (9a) - (9c) exists: namely $\mathbf{D}^0$.

Step 4 defines a flexibility unit price κ. The definition of a flexibility price is central to our proposed mechanism. Rather than enforce a specific price function, here we present properties that a price of flexibility should satisfy. Later in Section 4.3 we provide examples

[5] In the real-world scenario of renewable curtailment, $c_{\min} = 0$ (since marginal cost of renewables is taken to be 0) and $\mathcal{T}$ is simply the set of intervals for which renewables are curtailed.

that satisfy the given properties. Before introducing them, we first we define the concept of a *flexibility surplus*.

Definition 4.1. The flexibility surplus $S := \sum_t (\pi_t^0 - \tilde{\pi}_t)\mathbf{1}^\top \tilde{\mathbf{d}}_t$ *is the difference between the total demand-side energy payment if demand were paying baseline energy price* π_t^0 *and the total demand-side payment when demand pays the price,* $\tilde{\pi}_t$*. Because of Lemma B.2 (see Appendix B) and constraint (1c),* $S \geq 0$*. We interpret S as the improvement in welfare (over the baseline) of the demand side of the market as a whole when the dispatch* $\tilde{\mathbf{d}}_t$ *optimally utilizes flexibility.*

Now we establish properties that should be satisfied by a flexibility price κ:

- κ is uniform (each load faces the same κ)
- $\kappa_t \geq 0$ when $\sum_j \Delta_{j,t} \geq 0$ and $\kappa_t \leq 0$ when $\sum_j \Delta_{j,t} \leq 0$. This means that the payment for both up and down flexibility is non-negative, as at least some flexibility in both directions is necessary to dispatch shiftable demand.
- The sum of all flexibility payments equals the flexibility surplus: $\sum_t \kappa_t \mathbf{1}^\top(\tilde{\mathbf{d}}_t - \mathbf{d}_t^0) = S$.

This last property is natural, as our scheme distributes the surplus arising from the increased economic efficiency of the flexibility dispatch to the loads that provide this flexibility. Another desirable property we seek when constructing κ_t is that its magnitude should reflect the value of flexibility to the system in interval t.

Step 5 maximizes social welfare for the demand side of the market given flexibility price κ and energy price π^0. In order to construct this welfare maximization problem, we need to update the definition of demand utility with a term that quantifies the benefit that comes from offering flexibility.

Definition 4.2. Let $\mathbf{d}_j \in \mathbb{R}^T$ *be load j's consumption vector and* κ *and* π *be the flexibility and energy price vectors, respectively.* $\mathbf{d}_j^0$ *is the load's reported baseline. Then demand utility is given by*

$$u_j^*(\mathbf{d}_j; \pi, \kappa, \mathbf{d}_j^0) = U_j - \pi^\top \mathbf{d}_j + \kappa^\top(\mathbf{d}_j - \mathbf{d}_j^0)$$

Next, we solve a demand allocation optimization where the total demand dispatch amount in each interval is fixed to be equal to the total interim demand dispatch from Step 3. This allows the settlement for the generation side of the market to remain unaffected by the redistribution on the demand side.

$$\max_{\mathbf{d}_j\ \forall j} \quad \sum_j u_j^*(\mathbf{d}_j) \tag{11a}$$

subject to

$$\mathbf{1}^\top \mathbf{d}_j = E_j \qquad \forall j \tag{11b}$$

$$\underline{\mathbf{d}}_j \leq \mathbf{d}_j \leq \overline{\mathbf{d}}_j \qquad \forall j \tag{11c}$$

$$d_{j,t} \geq d_{j,t}^0 \qquad \forall j,\ \forall t \in \mathcal{T} \tag{11d}$$

$$d_{j,t} \leq d_{j,t}^0 \qquad \forall j,\ \forall t \in \mathcal{T}^c \tag{11e}$$

$$\mathbf{1}^\top \mathbf{d}_t = \mathbf{1}^\top \tilde{\mathbf{d}}_t \qquad \forall t \tag{11f}$$

The optimal solution of the above problem $\mathbf{d}_j{}^*$ determines the actual consumption of load j over the horizon.

Finally, Step 6 settles the market with $(\tilde{\mathbf{p}}_i\ \forall i, \tilde{\pi})$ for generators and $(\mathbf{d}_j^*\ \forall j, \pi^0, \kappa)$ for loads. Load j *pays* ${\pi^0}^\top \mathbf{d}_j^*$ for energy because it is the price it *would* have payed in the baseline scenario. The load *receives* $\kappa^\top \Delta_j^*$ for deviating by Δ_j^* from its baseline.

4.2 Analyzing participation incentives

The following theorem establishes properties for both generator and load utility under the proposed market mechanism and settlement scheme. We show that incentives are aligned on both sides of the market.

Theorem 4.3. *Let* $(\tilde{\mathbf{p}}_i, \mathbf{d}_j^*, \Delta_j^*)$ *be the energy and flexibility allocation from the market mechanism. Let* $(\tilde{\pi}, \pi^0, \kappa)$ *be the corresponding energy and flexibility prices. Then*

(i) $(\tilde{\mathbf{p}}_i, \mathbf{d}_j^*, \Delta_j^*)$ *clears the market;*

(ii) $(\tilde{\mathbf{p}}_i, \mathbf{d}_j^*, \Delta_j^*, \tilde{\pi}, \pi^0, \kappa)$ *is revenue neutral for the market operator;*

(iii) $(\tilde{\mathbf{p}}_i, \tilde{\pi})$ *provides dispatch-following incentives for generators and satisfies their participation constraints;*

(iv) $(\mathbf{d}_j^*, \Delta_j^*, \pi^0, \kappa)$ *satisfies participation constraints for loads. Specifically, for each j,*

$$u_j^*(\mathbf{d}_j^*; \pi^0, \kappa, \mathbf{d}_j^0) \geq u_j(\mathbf{d}_j^0; \pi^0);$$

(v) For j for which $\Delta_{j,t}^* = 0$ *for all t (no flexibility offered),*

$$u_j^*(\mathbf{d}_j^*; \pi^0, \kappa, \mathbf{d}_j^0) \leq u_j(\tilde{\mathbf{d}}_j; \tilde{\pi});$$

For j for which $|\Delta_{j,t}^*| > 0$ *for some t (flexibility offered),*

$$\sum_j u_j^*(\mathbf{d}_j^*; \pi^0, \kappa, \mathbf{d}_j^0) \geq \sum_j u_j(\tilde{\mathbf{d}}_j; \tilde{\pi}).$$

Statement (v) is of particular importance and highlights that loads are better off offering flexibility than not: no load becomes worse off than at its baseline consumption but loads that do offer flexibility are (weakly) better off as a group than those that do not.

Proof. See Appendix C. □

4.3 A price for flexibility

A core piece of our proposed market design is the flexibility price κ. How to properly compensate demand for flexibility is a challenging open question. Flexibility, as defined in this work, is a public good: in the interim, energy price-only settlement $(\tilde{\mathbf{P}}, \tilde{\mathbf{D}}, \tilde{\pi})$, even those loads who do not offer relaxed bounds on their consumption (i.e., offer flexibility to the market) still benefit from others that do by enjoying a lower price. To address this, our mechanism directly pays flexible loads that based on how much of the flexibility they offer is dispatched. We proceed in two stages: first we define a flexibility price κ that satisfies certain desirable properties (Step 4 in the mechanism); and second, we compute an allocation of energy and flexibility that maximizes individual utility while also respecting the previously-determined generation dispatch (Step 5 in the mechanism).

To this point, we have made the second stage concrete with (11a) - (11f) but we have not yet given specific examples of flexibility prices that satisfy the desirable properties of κ listed in Step 4 above. In this section we propose three different flexibility prices, commenting on their relative advantages. An interesting future research direction is to explore other forms of this price.

4.3.1 Optimization-based. Our first approachis to directly solve an optimization problem with the properties listed in Step 4 as constraints.

$$\begin{aligned} \min_{\kappa \in \mathbb{R}^T} \quad & f(\kappa) \\ \text{s.t.} \quad & S = \sum_t \kappa_t \mathbf{1}^\top (\tilde{\mathbf{d}}_t - \mathbf{d}_t^0) \\ & \kappa_t \geq 0 \qquad \forall t \in \mathcal{T} \\ & \kappa_t \leq 0 \qquad \forall t \in \mathcal{T}^{\mathrm{c}} \end{aligned}$$

A benefit of this approach is that the choice of an objective function $f(\kappa)$ can be made in order to enforce desired structural properties. For example, setting $f(\kappa) = ||\kappa||_2$ yields a smooth price schedule. If prices that weight high-value time-periods more are desired, then one could have $f(\kappa) = ||\kappa||_1$.[6]

The adaptability of this formulation of κ is its main advantage. One potential disadvantage is that it does not yield a closed-form representation of κ in general, which could make the price difficult to interpret. The subsequent two designs we consider provide closed-form representations of κ.

4.3.2 Budget-balance. A contrasting formulation of κ is based on the market operator's budget balance condition:

$$\sum_t \tilde{\pi}_t \mathbf{1}^\top \tilde{\mathbf{p}}_t = \sum_t \pi_t^0 \mathbf{1}^\top \mathbf{d}_t^* - \sum_t \kappa_t \mathbf{1}^\top \left(\mathbf{d}_t^* - \mathbf{d}_t^0 \right) \tag{12}$$

This condition states that the total payments to generators equals the total energy payments from demand minus the total flexibility payments to loads. Noting that $\mathbf{1}^\top \tilde{\mathbf{p}}_t = \mathbf{1}^\top \tilde{\mathbf{d}}_t = \mathbf{1}^\top \mathbf{d}_t^*$ and enforcing equality for each t separately, we solve for (12) for κ_t to get

$$\kappa_t := \frac{\left(\pi_t^0 - \tilde{\pi}_t\right) \mathbf{1}^\top \mathbf{d}_t^*}{\mathbf{1}^\top \left(\mathbf{d}_t^* - \mathbf{d}_t^0 \right)}.$$

This form of κ_t satisfies our desired properties but it does have an important drawback. When $t \in \mathcal{T}$, $\tilde{\pi}_t = \pi_t^0 = c_{\min}$, which implies that $\kappa_t = 0$. So, κ_t is never strictly positive, which leads to only down-flexibility ($\Delta_t < 0$) being rewarded. Further, for $t \in \mathcal{T}$, $\kappa_t = 0$ and so the prices do not capture the time-varying value of flexibility for those intervals. The following design avoids these disadvantages.

4.3.3 Flexibility surplus. Another closed-form version of κ_t can be defined using the flexibility surplus:

$$\kappa_t = \begin{cases} \frac{S}{2} \frac{P_t^{\mathrm{cap}} - \mathbf{1}^\top \mathbf{d}_t^0}{\sum_t P_t^{\mathrm{cap}} - \mathbf{1}^\top \mathbf{d}_t^0} \frac{1}{\mathbf{1}^\top \left(\tilde{\mathbf{d}}_t - \mathbf{d}_t^0 \right)}, & t \in \mathcal{T} \\ \frac{S}{2} \frac{\tilde{\pi}_t}{\sum_t \tilde{\pi}_t} \frac{1}{\mathbf{1}^\top \left(\tilde{\mathbf{d}}_t \mathbf{d}_t^0 \right)}, & t \in \mathcal{T}^{\mathrm{c}} \end{cases}$$

This form of κ_t is the product of three terms in both cases. The first, $\frac{S}{2}$, divides the total flexibility surplus evenly between up- and down-flexibility periods. The second term distributes that half-surplus amongst the time intervals. For $t \in \mathcal{T}$, an interval receives an amount proportional to its surplus (i.e., curtailed) lowest-cost generation. For $t \in \mathcal{T}^{\mathrm{c}}$, an interval receives the amount proportional to the interim price $\tilde{\pi}_t$in that time period. The third term divides by the total allocation of flexibility, as determined by the dispatch from (1a) - (1e) with (9a) - (9c).

Like the previous two flexibility prices, this κ satisfies all of the desired properties including budget balance. Its two-part specification reflects the different function flexibility has in $\mathcal{T}$ versus $\mathcal{T}^{\mathrm{c}}$. In $\mathcal{T}$, flexibility allows otherwise-curtailed low-cost generation to be dispatched. In $\mathcal{T}^{\mathrm{c}}$, flexibility allows for lesser amounts of more-costly generation to be dispatched. While this formulation addresses the zero-price shortcoming of the budget-balance formulation and has a closed form representation, it is vulnerable to volatility when $\mathbf{1}^\top \mathbf{d}_t^* - \mathbf{1}^\top \mathbf{d}_t^0$ is small.

Comparing these three formulations, it is worth highlighting that, while a closed-form κ might be desirable for reasons convenience and interpretability, the optimized-based approach is more principled and adaptable. For this reason we choose to implement that version of the flexibility price in the case study in the next section.

5 CASE STUDY

We conclude the paper with a demonstration of the our new market dispatch of flexibility on a test case derived from the real-world CAISO market. Our numerical results show a significant increase in utility for loads when they allow their flexibility to be dispatched by the market operator, thus highlighting the value of redesigning the market to ensure participation incentives of loads are aligned.

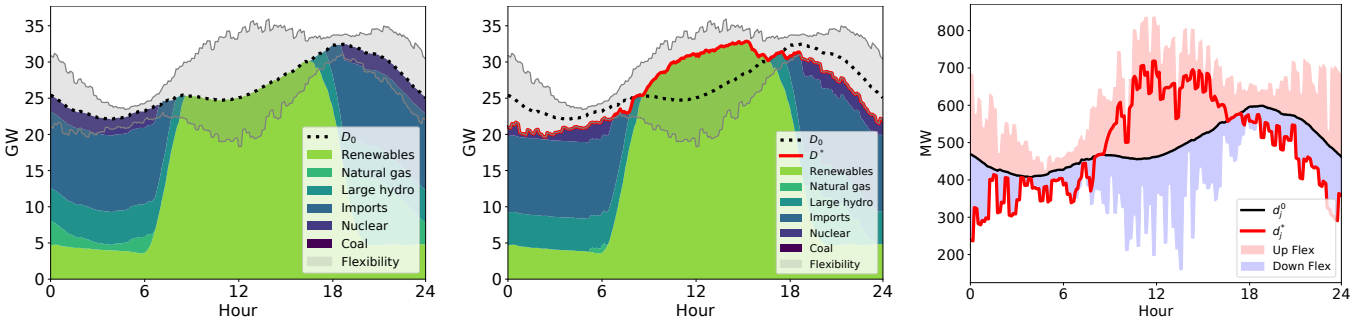
5.1 Setup

Disaggregated demand-side data for bulk electricity markets is not readily available [23]. We therefore take existing publicly-available generation and aggregate load data from CAISO and simulate a demand side to to the market. Our simulations are implemented in Python and all optimization problems are solved with `CVXPY`[5, 20]. The simulation were run on a 2019 MacbookPro (2.8 GHz Quad-Core i7, 16GB RAM).

Throughout our experiments, we ran the single-shot market mechanism described in Section 4, which assumes an accurate demand forecast, and computed the flexibility price κ using the optimization-based formulation.

The test cases are constructed as follows. Generation time series data, disaggregated by resource type (e.g., renewable, hydro, coal), from July 2, 2020 is obtained from [11]. The data have observations every 5 minutes for 24 hours (288 total). At their peak, renewables (e.g., wind, solar, small hydro, biomass) account for approx. 60% of the net generation. We clean the data by removing trivial generation resources like batteries and negative values for solar generation at night (due to concentrating solar); the result is 6 generation resource types: renewables, natural gas, large hydro, nuclear, coal, and external imports from adjacent control areas. The aggregate demand D_0 profile is obtained from the resulting net generation. We scale up the entire renewable profile by 220% so that there is a set of intervals $\mathcal{T}$ where renewable generation alone exceeds aggregate demand and thus renewables must be curtailed. As we noted previously, this scenario is not (yet) the case in California but in other markets has already begun occurring [3].

[6] If $f(\kappa)$ is a norm, then this formulation has the additional property that $\kappa_t = 0$ only if $\mathbf{1}^\top (\tilde{\mathbf{d}}_t - \mathbf{d}_t^0) = 0$. This means that intervals that do not dispatch flexibility will not have a non-zero price.



(a) Baseline dispatch (demand does not offer flexibility)

(b) Flexibile dispatch (demand offers flexibility)

(c) Individual load profile for $j = 29$: baseline (black) and flexible dispatch (red)

Figure 2: Comparison of the baseline market with the proposed market design in a CAISO case study.

We assume that conventional generation types and imports are dispatchable up and down without ramping limits, whereas renewables can only be curtailed. We also make the simplifying assumption that conventional generation can produce any amount from 0 to their upper limits, which are taken from the original data to be the maximum production at any point in the 24 hour window. Unit cost data in $/MWh are the Variable O&M costs for 2020 from EIA's Annual Energy Outlook, see Table 1 in [2]. Unit costs for imports were assumed to be the average of costs for the other generation types present in our simulation.

The aggregate demand profile from the CAISO data (black dotted line in Figure 2a) is disaggregated proportionally into individual load profiles. These profiles are then perturbed with random noise to introduce temporal variability to the relative fraction of the aggregate each individual load consumes. The number of individual loads m can be set arbitrarily and in our case study here, $m = 30$.[7] Half of these were designated inflexible loads and the other half to flexible loads. Centered around each of the individual flexible load profiles are upper and lower bound profiles for the consumption of each load in each time interval. These bounds are generated with a sinusoidal function which allows parametric scaling of flexibility by varying the amplitude and phase. S We note here that despite not being able to access real-world load profiles, our load disaggregation scheme produces qualitatively similar results to the load shapes in [47]. The baseline load profile (black) and the flexibility range (grey) are shown for the market in aggregate in Figure 2a and for an individual load in Figure 2c.

5.2 Results

Figure 2 provides a detailed contrast between the traditional baseline market design, under which shiftable demands do not offer their flexibility, and the proposed design of this paper, under which shiftable demands have incentives to expose their flexibility. The reduction of curtailment of renewable generation that results from shiftable demands is immediately clear from these figures.

In more detail, Figure 2a shows the generation dispatch as well as the baseline aggregate demand. The available aggregate flexibility is shown as a light grey overlay. Notice that renewables are curtailed between hours 8 and 17, as there is an excess supply available to meet the baseline aggregate demand D_0.

[7] Experiments with other values of m did not change results qualitatively.

Figure 2a should be contrasted with Figure 2b, which shows the market dispatch (aggregate shown in red) when flexible demand is utilized. The flexibility upper/lower bounds (grey) and the baseline aggregate demand (dotted black) are superimposed for comparison. Load is dispatched *up* in periods with curtailed renewable (hours 8 - 17) and dispatched *down* during the remaining hours to compensate. In this simulation, for the hours when load is dispatched down, the lower bound on flexibility is often tight whereas the upper bound is not attained at any point over the time horizon. This highlights the point that both up- and down-flexibility are required in equal amounts due to the equality constraint for total demand over the time horizon (e,g., (1c), 11b). The limiting factor for shiftable loads to increase demand during the middle of the day (and therefore reduce renewable curtailment) could actually be their inability to reduce its demand at other times.

Figure 2c drills deeper and considers the profile of an individual load. This figure shows that feasibility of the flexible market dispatch for the load is indeed satisfied, as required by the constraints in (11b) - (11e). The black curve shows the baseline demand $\mathbf{d}_j^0$ for load $j = 29$ and the red curve shows the dispatch with flexibility $\mathbf{d}_j^*$. Both trajectories respect the upper and lower flexibility bounds $\overline{\mathbf{d}}_j, \underline{\mathbf{d}}_j$. Further, all loads (and therefore the aggregate load as well) change their dispatch under the flexibility dispatch allocation in the same direction (i.e., up or down) in each interval. This is due to constraints (9b) and (9c), without which the undesirable scenario where some loads increase and other simultaneously decrease their consumption could occur.

The case study also provides a concrete illustration of many of the properties of prices we proved previously. In particular, the top panel of Figure 3 shows this graphically that the baseline price π^0 is a lower bound for $\tilde{\pi}$, a property proven in Lemma B.2. The lower panel of Figure 3 illustrates that the flexibility price κ satisfies its desired properties in that it is positive when up-flexibility is dispatched and negative when down-flexibility is dispatched. Its magnitude also reflects a time-varying value of flexibility; specifically, κ_t is most positive during the middle of the day when renewables have peak capacity and load should be dispatched up to utilize them and is most negative early and late in the day when expensive conventional generation dominates the generation mix and load should be dispatched down to reduce cost.

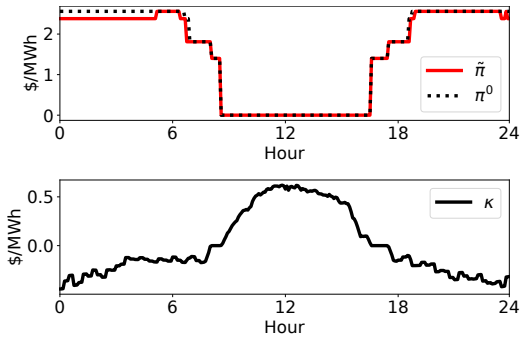


Figure 3: Illustration of energy prices $\tilde{\pi}$ and π^0 (top) and flexibility price κ (bottom) in the CAISO case study.

Table 1: Total revenue, cost, and utility for generation and demand in CAISO case study. Amounts are in millions.

	Baseline	With Flexibility	% change
Total Generation Revenue	\$11.91	\$10.58	-11.14
Total Generation Cost	\$6.39	\$5.39	-15.62
Total Generation Utility (profit)	\$5.52	\$5.19	-5.95
Total Demand Cost	\$11.91	\$10.90	-8.49
Total Demand Utility	\$-11.91	\$-10.58	+11.14
Total Flexibility Payment	\$0.00	\$0.32	—

In Table 1 we quantify the economic value of the proposed market design as compared to the baseline design by comparing market participant utility gains/losses between the two scenarios. The first observation from this table is that the demand side of the market increases its utility by 11% over the baseline while only needing to re-dispatch 10% of its total load. As flexibility is provided by the demand side of the of the market, our mechanism increases their utility to compensate.

The second observation is that each load individually is at least as well off under the flexibility mechanism as under the baseline scenario, but loads that offer flexibility are better off than those that do not. This can be seen by comparing total demand cost of \$10.90 to the total demand utility of -\$10.58. The difference in the magnitudes of these values is exactly the flexibility payment of \$0.32. Inflexible loads pay for energy but do not receive any benefit from the flexibility payment, which only goes to flexible loads.

Third is that generators are worse off under the flexibility mechanism due to a lower energy price $\tilde{\pi}$. Dispatching flexibility improves the overall efficiency (i.e., generation cost) of the dispatch but because the spot price decreases as well, that benefit is not captured by generators, instead going to the loads. From a generator's point of view, this is not desirable as it will lower their profits individually and collectively. However we remark that *any* improvement in market efficiency is likely to lower generator profits (for additional discussion of these see [33]). That does not mean that improvements in market efficiency ought to be avoided though. Rather, we take the view that incentives for improving system efficiency should be aligned with those of the market participants who actually provide the efficiency-improving service. In the setting we explore in this work, the deserving participants are flexible loads with shiftable demand.

6 CONCLUSION

This paper focuses on a crucial and under-explored aspect of demand response markets: the incentives of loads with shiftable demand to expose flexibility to the market operator. We first show that relying on the energy spot price alone to compensate loads—as the standard market design does—leads to incentive misalignment: demand might end up worse off bidding flexibly than inflexibly. Our market mechanism addresses this shortcoming in two parts. The first constrains the total amount of flexibility that can be dispatched in each period, ensuring that costly generators cannot be dispatched. The second introduces a flexibility price and distributes the surplus that arises from the more efficient dispatch to loads that offer flexibility.

The flexibility price serves two useful purposes. One is to provide a time-varying signal to loads about the most profitable times to offer their flexibility to the market. A second value of the flexibility is to correct a free-rider problem that arises in an energy price-only market: flexibility is a public good, which means that all loads benefit from flexibility whether they contribute it themselves or not. In our mechanism, the flexibility payment, which is the product of flexibility price and flexibility dispatch, is only non-zero for flexible loads.

Importantly, our proposed mechanism has the same basic structure as the current economic dispatch market design, which provides a pathway to adoption. In this work though, our model sets aside several real-world electricity market features like startup costs, ramping constraints, line congestion, and rolling window market clearing. These undeniably impact market dispatch and are the focus of large portion of research on electricity market design. However they are typically evaluated without considering a responsive demand side of the market. In contrast, our focus here is on the mechanism for incorporating shiftable demand into the economic dispatch framework and analyzing the incentives that result. It will be important in future work to tease out how the above-mentioned generation-side characteristics interact with the demand-side structure in our model.

Finally, another important open problem motivated by our work relates to flexibility pricing. Our market design shows how to incorporate a flexibility price into the marketplace and proposes three potential designs for flexibility prices. The flexibility prices we introduce satisfy the minimal desired properties, but each have some drawbacks and thus a further exploration of the design of flexibility prices is an important research question. In particular, is there a stable and interpretable flexibility price, aligned both with individual and social welfare objectives, that incentivizes loads to bid their flexibility into the market?

ACKNOWLEDGMENTS

This work was supported by VMWare and NSF Grants CNS-2105648, AitF-1637598, CNS-1518941, NSF ECCS 1931662, and by the Resnick Sustainability Institute at Caltech. We thank the anonymous reviewers for their careful reading of this paper and for their insightful suggestions.

REFERENCES

[1] 2020. *California State Senate Bill 100: 100 Percent Clean Energy.* https://focus.senate.ca.gov/sb100

[2] 2021. *Cost and Performance Characteristics of New Generating Technologies, Annual Energy Outlook 2021.* Technical Report. Energy Information Administration (EIA).

[3] 2021. *Quarterly Energy Dynamics Q4 2020.* Technical Report. Australian Energy Market Operator (AEMO).

[4] HA Aalami, M Parsa Moghaddam, and GR Yousefi. 2010. Demand response modeling considering interruptible/curtailable loads and capacity market programs. *Applied Energy* 87, 1 (2010), 243–250.

[5] Akshay Agrawal, Robin Verschueren, Steven Diamond, and Stephen Boyd. 2018. A rewriting system for convex optimization problems. *Journal of Control and Decision* 5, 1 (2018), 42–60.

[6] Hunt Allcott. 2009. Real time pricing and electricity markets. *Harvard University* 7 (2009).

[7] Peter Alstone. 2017. *2025 California Demand Response Potential Study.* Technical Report LBNL-2001113. Lawrence Berkeley National Laboratory (LBNL).

[8] Navid Azizan, Yu Su, Krishnamurthy Dvijotham, and Adam Wierman. 2020. Optimal pricing in markets with nonconvex costs. *Operations Research* 68, 2 (2020), 480–496.

[9] Dimitri P Bertsekas. 2009. *Convex optimization theory.* Athena Scientific Belmont.

[10] Suzhi Bi and Ying Jun Zhang. 2013. False-data injection attack to control real-time price in electricity market. In *2013 IEEE Global Communications Conference (GLOBECOM)*. 772–777.

[11] CAISO. 2020. *California ISO Open Access Same-time Information System (OASIS).* Retrieved July 2, 2020 from http://oasis.caiso.com/mrioasis/logon.do

[12] California Independent System Operator (CAISO). 2017. *California ISO records banner year for renewables integration.* http://www.caiso.com/Documents/CaliforniaISORecordsBannerYearforRenewablesIntegration.pdf

[13] Yanxin Chai, Yue Xiang, Junyong Liu, Chenghong Gu, Wentao Zhang, and Weiting Xu. 2019. Incentive-based demand response model for maximizing benefits of electricity retailers. *Journal of Modern Power Systems and Clean Energy* 7, 6 (2019), 1644–1650.

[14] Hung-po Chao. 2011. Demand response in wholesale electricity markets: the choice of customer baseline. *Journal of Regulatory Economics* 39, 1 (2011), 68–88.

[15] Hung-po Chao, Stephen Peck, Shmuel Oren, and Robert Wilson. 2000. Flow-based transmission rights and congestion management. *The Electricity Journal* 13, 8 (2000), 38–58.

[16] Yongbao Chen, Peng Xu, Yiyi Chu, Weilin Li, Yuntao Wu, Lizhou Ni, Yi Bao, and Kun Wang. 2017. Short-term electrical load forecasting using the Support Vector Regression (SVR) model to calculate the demand response baseline for office buildings. *Applied Energy* 195 (2017), 659–670.

[17] Antonio J Conejo and Ramteen Sioshansi. 2018. Rethinking restructured electricity market design: Lessons learned and future needs. *International Journal of Electrical Power & Energy Systems* 98 (2018), 520–530.

[18] S Datchanamoorthy, S Kumar, Y Ozturk, and G Lee. 2011. Optimal time-of-use pricing for residential load control. In *2011 IEEE International Conference on Smart Grid Communications (SmartGridComm)*. IEEE, 375–380.

[19] Ruilong Deng, Zaiyue Yang, Mo-Yuen Chow, and Jiming Chen. 2015. A survey on demand response in smart grids: Mathematical models and approaches. *IEEE Transactions on Industrial Informatics* 11, 3 (2015), 570–582.

[20] Steven Diamond and Stephen Boyd. 2016. CVXPY: A Python-embedded modeling language for convex optimization. *Journal of Machine Learning Research* 17, 83 (2016), 1–5.

[21] Torgeir Ericson. 2009. Direct load control of residential water heaters. *Energy Policy* 37, 9 (2009), 3502–3512.

[22] Ben Foster. 2019. *2019 Assessment of Demand Response and Advanced Metering.* Technical Report. Federal Energy Regulatory Commission (FERC).

[23] Natalie Mims Frick, Eric Wilson, Janet Reyna, Andrew Parker, Elaina Present, Janghyun Kim, Tianzhen Hong, Han Li, and Tom Eckman. 2019. End-Use Load Profiles for the US Building Stock: Market Needs, Use Cases, and Data Gaps. (2019).

[24] Paul R Gribik, William W Hogan, Susan L Pope, et al. 2007. Market-clearing electricity prices and energy uplift. *Cambridge, MA* (2007).

[25] Theodore Groves and John O Ledyard. 1985. *Incentive compatibility ten years later.* Technical Report 648. Discussion Papers from Northwestern University, Center for Mathematical Studies in Economics and Management Science.

[26] Ye Guo, Cong Chen, and Lang Tong. 2021. Pricing multi-interval dispatch under uncertainty part I: Dispatch-following incentives. *IEEE Transactions on Power Systems* (2021).

[27] Karen Herter. 2007. Residential implementation of critical-peak pricing of electricity. *Energy policy* 35, 4 (2007), 2121–2130.

[28] William W Hogan. 1992. Contract networks for electric power transmission. *Journal of regulatory economics* 4, 3 (1992), 211–242.

[29] William W Hogan. 2020. Electricity Market Design: Multi-Interval Pricing Models. https://scholar.harvard.edu/files/whogan/files/hogan_yesenergy_202720r.pdf

[30] William W Hogan and Brendan J Ring. 2003. On minimum-uplift pricing for electricity markets. *Electricity Policy Group* (2003), 1–30.

[31] Bowen Hua, Dane A Schiro, Tongxin Zheng, Ross Baldick, and Eugene Litvinov. 2019. Pricing in multi-interval real-time markets. *IEEE Transactions on Power systems* 34, 4 (2019), 2696–2705.

[32] A Rezaee Jordehi. 2019. Optimisation of demand response in electric power systems, a review. *Renewable and sustainable energy reviews* 103 (2019), 308–319.

[33] Daniel S Kirschen. 2003. Demand-side view of electricity markets. *IEEE Transactions on power systems* 18, 2 (2003), 520–527.

[34] John O Ledyard. 1989. Incentive compatibility. In *Allocation, Information and Markets*. Springer, 141–151.

[35] Yen-Yu Lee, Jin Hur, Ross Baldick, and Salvador Pineda. 2010. New indices of market power in transmission-constrained electricity markets. *IEEE Transactions on Power Systems* 26, 2 (2010), 681–689.

[36] George Liberopoulos and Panagiotis Andrianesis. 2016. Critical review of pricing schemes in markets with non-convex costs. *Operations Research* 64, 1 (2016), 17–31.

[37] Yanchao Liu, Jesse T Holzer, and Michael C Ferris. 2015. Extending the bidding format to promote demand response. *Energy Policy* 86 (2015), 82–92.

[38] Peter B Luh, William E Blankson, Ying Chen, Joseph H Yan, Gary A Stern, Shi-Chung Chang, and Feng Zhao. 2006. Payment cost minimization auction for deregulated electricity markets using surrogate optimization. *IEEE Transactions on Power systems* 21, 2 (2006), 568–578.

[39] Johanna L Mathieu, Duncan S Callaway, and Sila Kiliccote. 2011. Examining uncertainty in demand response baseline models and variability in automated responses to dynamic pricing. In *2011 50th IEEE Conference on Decision and Control and European Control Conference*. IEEE, 4332–4339.

[40] Deepan Muthirayan, Enrique Baeyens, Pratyush Chakraborty, Kameshwar Poolla, and Pramod P Khargonekar. 2019. A minimal incentive-based demand response program with self reported baseline mechanism. *IEEE Transactions on Smart Grid* 11, 3 (2019), 2195–2207.

[41] Deepan Muthirayan, Dileep Kalathil, Kameshwar Poolla, and Pravin Varaiya. 2019. Mechanism design for demand response programs. *IEEE Transactions on Smart Grid* 11, 1 (2019), 61–73.

[42] Sepehr Ramyar, Andrew Lu Liu, and Yihsu Chen. 2019. Power market model in presence of strategic prosumers. *IEEE Transactions on Power Systems* 35, 2 (2019), 898–908.

[43] Navid Azizan Ruhi, Krishnamurthy Dvijotham, Niangjun Chen, and Adam Wierman. 2017. Opportunities for price manipulation by aggregators in electricity markets. *IEEE Transactions on Smart Grid* 9, 6 (2017), 5687–5698.

[44] Fred C Schweppe, Michael C Caramanis, Richard D Tabors, and Roger E Bohn. 2013. *Spot pricing of electricity.* Springer Science & Business Media.

[45] Pierluigi Siano. 2014. Demand response and smart grids—A survey. *Renewable and sustainable energy reviews* 30 (2014), 461–478.

[46] Fei Wang, Kangping Li, Chun Liu, Zengqiang Mi, Miadreza Shafie-Khah, and João PS Catalão. 2018. Synchronous pattern matching principle-based residential demand response baseline estimation: Mechanism analysis and approach description. *IEEE Transactions on Smart Grid* 9, 6 (2018), 6972–6985.

[47] Eric Wilson. 2019. *End-Use Load Profiles for the U.S. Building Stock.* Technical Report. National Renewable Energy Laboratory (NREL).

[48] Jinye Zhao, Tongxin Zheng, and Eugene Litvinov. 2019. A multi-period market design for markets with intertemporal constraints. *IEEE Transactions on Power Systems* 35, 4 (2019), 3015–3025.

A PROOF OF THEOREM 3.2

Start by forming the Lagrangian for (1a) - (1e):

$$\begin{aligned}\mathscr{L}(\mathbf{p}_i, \mathbf{d}_j; \lambda, \rho_j, \mu_i^{\pm}, \eta_j^{\pm}) = &\sum_i c_i(\mathbf{p}_i) \\ &+ \sum_t \lambda_t(\mathbf{1}^\top \mathbf{d}_t - \mathbf{1}^\top \mathbf{p}_t) + \sum_j \rho_j \mathbf{1}^\top \mathbf{d}_j \\ &+ \mu_i^{+\top}(\mathbf{p}_i - \overline{\mathbf{p}}_i) - \mu_i^{-\top}(\mathbf{p}_i - \underline{\mathbf{p}}_i) \\ &+ \eta_j^{+\top}(\mathbf{d}_j - \overline{\mathbf{d}}_j) - \eta_j^{-\top}(\mathbf{d}_j - \underline{\mathbf{d}}_j)\end{aligned}$$

We assume that (1a) - (1e) has a feasible point. Let $(\mathbf{p}_i^*, \mathbf{d}_j^*, \lambda^*, \rho_j^*, \mu_i^{\pm *}, \eta_i^{\pm *})$ denote its optimal solution, which exists because the c_i are continuous and the feasible set is compact. Compactness follows from constraint (1d) where it can be seen that all entries of $\mathbf{p}_i^*$ and $\mathbf{d}_j^*$ must be finite. Strong duality holds because all constraints are affine and the objective function is convex (see e.g., Prop. 5.3.1 in [9]).

Therefore the following KKT stationarity condition for holds for every i:

$$\frac{\partial \mathscr{L}}{\partial \mathbf{p}_i}(\mathbf{p}_i^*) = \nabla c_i(\mathbf{p}_i^*) - \lambda^* + \mu_i^{+*} - \mu_i^{-*} = \mathbf{0} \tag{13}$$

We compute the derivative of the λ term by noting that

$$\frac{\partial}{\partial p_{i,t}} \sum_t \lambda_t(\mathbf{1}^\top \mathbf{p}_t - \mathbf{1}^\top \mathbf{d}_t) = \lambda_t$$

Stacking this equation for each t into vector form gives

$$\frac{\partial}{\partial \mathbf{p}_i} \sum_t \lambda_t(\mathbf{1}^\top \mathbf{p}_t - \mathbf{1}^\top \mathbf{d}_t) = \lambda.$$

The price vector, as defined in (2), is $\pi = \lambda^*$. Thus (13) is

$$\nabla c_i(\mathbf{p}_i^*) - \pi + \mu_i^{+*} - \mu_i^{-*} = \mathbf{0} \tag{14}$$

Next, we rewrite (4) equivalently as a minimization of $-u_i$ over the same feasible set and take its Lagrangian.[8]

$$\mathscr{L}_i(\mathbf{p}_i; \mu_i^{\pm}) = c_i(\mathbf{p}_i) - \pi^\top \mathbf{p}_i + \mu_i^{+\top}(\mathbf{p}_i - \overline{\mathbf{p}}_i) - \mu_i^{-\top}(\mathbf{p}_i - \overline{\mathbf{p}}_i)$$

The KKT stationarity condition is

$$\frac{\partial \mathscr{L}_i}{\partial \mathbf{p}_i} = \nabla c_i(\mathbf{p}_i) - \pi + \mu_i^{+} - \mu_i^{-} = \mathbf{0} \tag{15}$$

It is clear that $\mathbf{p}_i = \mathbf{p}_i^*, \pi = \lambda^*, \mu_i^{\pm} = \mu_i^{\pm *}$ is a solution to (15) because $(\mathbf{p}_i^*, \lambda^*, \mu_i^{\pm *})$ satisfies (14).

Now we show that $\mathbf{p}_i^*$ satisfies participation constraints. Outside of the mechanism, the generator would produce $\mathbf{p}_i = \mathbf{0}$ with $u_i(\mathbf{0}) = 0$. This is because we assumed that $c_i(\mathbf{0}) = 0$. We need to show that 0 is a lower bound for $u_i(\mathbf{p}_i^*)$.

In (14), when ${\mu_{i,t}^+}^* > 0$ then ${\mu_{i,t}^-}^* = 0$ as only one of the lower/upper bounds can be attained at a time. But if ${\mu_{i,t}^-}^* > 0$ then $p_{i,t}^* = 0$. When ${\mu_{i,t}^+}^* = {\mu_{i,t}^-}^* = 0$, then $p_{i,t}^* > 0$ and $\frac{\partial c_{i,t}}{\partial p_{i,t}}(p_{i,t}^*) = \pi_t$. Therefore $\pi_t p_{i,t}^* - c_{i,t}(p_{i,t}^*) = 0$. Finally, when ${\mu_{i,t}^+}^* > 0$, then ${\mu_{i,t}^-}^* = 0$ and $\frac{\partial c_{i,t}}{\partial p_{i,t}}(p_{i,t}^*) < \pi_t$. So $\pi_t p_{i,t}^* - c_{i,t}(p_{i,t}^*) > 0$. In each of these three situations we get that $u_i(\mathbf{p}_i^*) \geq 0$.

[8]Note that we use the same names for primal/dual variables in the individual problem as in the market dispatch problem. Although these variables *do not* represent the same quantities, we hope this is not cause for confusion.

B LEMMAS FOR THEOREM 4.3

This section contains two lemmas used in the proof of Theorem 4.3. Let $\mathbf{d}^0$ and $\mathbf{p}^0$ be the optimal primal solutions to (1a) - (1e) in the baseline case (i.e., $\underline{\mathbf{d}} = \overline{\mathbf{d}}$), and let λ^0 be the associated optimal dual variable for (1b). In the presence of flexible demand, that is, $\underline{d}_{j,t} < \overline{d}_{j,t}$ for some t and j, we get the following result.

Lemma B.1. *Let $\tilde{\mathbf{d}}_j$ for all j be optimal solutions of (1a) - (1e) with added constraints (9a - 9c). Then for all j*

$$u_j(\tilde{\mathbf{d}}_j) \geq u_j(\mathbf{d}_j^0).$$

The proof of this claim requires another technical lemma, which we state and prove below before returning to the proof of Lemma B.1. To proceed, we associate dual variables $\beta \in \mathbb{R}^{|\mathcal{T}|}$, $\gamma^+ \in \mathbb{R}^{|\mathcal{T}|}$, $\gamma^- \in \mathbb{R}^{|\mathcal{T}^c|}$ with constraints (9a) - (9c) respectively.

Lemma B.2. *$\lambda_t^0 \geq \tilde{\lambda}_t$ for $t \in \mathcal{T}^c$.*

Proof of Lemma B.2. Let $(\tilde{\mathbf{P}}, \tilde{\mathbf{D}}, \tilde{\lambda}, \tilde{\rho}, \tilde{\mu}^{\pm}, \tilde{\eta}^{\pm}, \tilde{\beta}, \tilde{\gamma}^{\pm})$ be the optimal primal/dual solution of (1a) - (1e) with added constraints (9a) - (9c). The arguments for the existence of this solution and the existence of strong duality are the same as those given in the proof of Theorem 3.2 (see Appendix A).[9]

For each t, there is a set of marginal generators $\mathcal{N}_t \subseteq \{1, \ldots, n\}$. By its definition, a marginal unit produces strictly between its upper and lower bounds. Therefore, for $i \in \mathcal{N}_t$

$$\tilde{\mu}_{i,t}^{+} = \tilde{\mu}_{i,t}^{-} = \mathbf{0}.$$

The KKT stationarity condition w.r.t. $\mathbf{p}_t$ is

$$\frac{\partial \mathscr{L}}{\partial \mathbf{p}_t}(\tilde{\mathbf{p}}_t) = \nabla c(\tilde{\mathbf{p}}_t) - \tilde{\lambda}_t \mathbf{1} + \tilde{\mu}_t^{+} - \tilde{\mu}_t^{-} = \mathbf{0}. \tag{16}$$

From this equation we have that **for all** $i \in \mathcal{N}_t$,

$$\tilde{\lambda}_t = \frac{\partial c_{i,t}}{\partial p_{i,t}}\left(\tilde{p}_{i,t}\right). \tag{17}$$

That is, all marginal costs are equal in that time for the marginal units.

Next, we claim that no generator produces more than its baseline; that is $\tilde{p}_{i,t} \leq p_{i,t}^0$ for all $i \in \{1, \ldots, n\}$. In the baseline scenario there are three groups of generators: those producing at their upper bound, those producing at their lower bound, and the marginal units. Those already at their upper bound are unable to increase their production. Increasing the production of a unit at its lower bound would incur a higher cost than increasing production by the same amount for a marginal unit. In (17) we argued that all marginal units have the same marginal cost at the optimal point. Therefore they all would increase production or all decrease. Due to the convexity and monotonicity of the cost functions, a decrease in production would result in a lower value for λ_t by (17).

This leaves two remaining possibilities: 1) at least one generator at its upper bound in the baseline case decreases its production or 2) a marginal unit decreases its production. In the first case, λ_t is unaffected since its value is determined by the cost function of a marginal unit. In the second case, due to the convexity and

[9]We assumed previously that a baseline solution exists for (1a) - (1e). A feasible point for (1a - 1e) with added constraints (9a - 9c) is just this baseline solution.

monotonicity of the cost functions, a decrease in production would result in a smaller value of λ_t by (17).

For all $t \in \mathcal{T}^c$, $\tilde{d}_{j,t} \leq d^0_{j,t}$ for each j by constraints (9c). Thus we have $\mathbf{1}^\top \tilde{\mathbf{d}}_t \leq \mathbf{1}^\top \mathbf{d}^0_t$ for $t \in \mathcal{T}^c$. By the power balance constraint (1b),

$$\mathbf{1}^\top \tilde{\mathbf{p}}_t \leq \mathbf{1}^\top \mathbf{p}^0_t.$$

From this we conclude that for at least one $i \in \mathcal{N}_t$, $p^0_{i,t} \geq \tilde{p}_{i,t}$. Since we take the cost functions to be convex and if $p^0_{i,t} \geq \tilde{p}_{i,t}\ \forall i$ as we showed just above, then

$$\frac{\partial c_{i,t}}{\partial p_{i,t}}(p^0_{i,t}) \geq \frac{\partial c_{i,t}}{\partial p_{i,t}}(\tilde{p}_{i,t}) \tag{18}$$

We conclude that

$$\lambda^0_t \geq \tilde{\lambda}_t.$$

□

Proof of Lemma B.1. We take the aggregate marginal cost curve of (1a) - (1e) to be left continuous. This is equivalent to always picking the smallest value of the subgradient of $\sum_i c_i(\mathbf{p}_i)$ in the KKT condition for $\mathbf{p}_i$ when the subgradient is not unique.

For $t \in \mathcal{T}$, the following are true:

- $\tilde{d}_{j,t} \geq d^0_{j,t}$ for all j by primal feasibility from constraint (9b);
- $\lambda^0_t = \tilde{\lambda}_t = c_{\min}$ because of (9c). By definition of $\mathcal{T}$, $c_{\min}$ is always the marginal cost in $\mathcal{T}$. Note that this claim requires the assumption from the beginning of the proof. Otherwise, when constraint (9a) is tight, it could occur that $\tilde{\lambda}_t > \lambda^0_t$.

Analogously for $t \in \mathcal{T}^c$ we have:

- $\tilde{d}_{j,t} \leq d^0_{j,t}$ for all j by primal feasibility from constraint (9c);
- $\tilde{\lambda}_t \leq \lambda^0_t$ by Lemma B.2.

By definition of the price π, $\pi^0 \geq \tilde{\lambda}$ follows from the above. We have also established that $\tilde{d}_{j,t} \geq d^0_{j,t}$ only when $\tilde{\lambda}_t = \lambda^0_t = c_{\min}$. Otherwise, $d_{j,t} \leq d^0_{j,t}$.

By definition of load utility in (5), we have that

$$\begin{aligned} u_j(\tilde{\mathbf{d}}_j) &= U_j - \sum_{t\in\mathcal{T}} \tilde{\pi}_t \tilde{d}_{j,t} - \sum_{t\in\mathcal{T}^c} \tilde{\pi}_t \tilde{d}_{j,t} \\ &= U_j - \sum_{t\in\mathcal{T}} c_{\min} \tilde{d}_{j,t} - \sum_{t\in\mathcal{T}^c} \tilde{\pi}_t \tilde{d}_{j,t} \\ &\geq U_j - \sum_{t\in\mathcal{T}} c_{\min} d^0_{j,t} - \sum_{t\in\mathcal{T}^c} \pi^0_t d^0_{j,t} \\ &= u_j(\mathbf{d}^0_j). \end{aligned}$$

□

C PROOF OF THEOREM 4.3

We prove each statement from the theorem in order, making use of the technical lemmas in the previous section.

(i) First note that $\tilde{\mathbf{D}}$ is a feasible solution for (11b) - (11f). The optimal solution $\mathbf{D}^*$ of (11a) - (11f) satisfies $\mathbf{1}^\top \mathbf{d}^*_t = \mathbf{1}^\top \tilde{\mathbf{d}}_t$ for all t due to primal feasibility.[10] Similarly, the generation dispatch $\tilde{\mathbf{P}}$ satisfies $\mathbf{1}^\top \tilde{\mathbf{p}}_t = \mathbf{1}^\top \tilde{\mathbf{d}}_t$ for all t by constraint (1b) and primal feasibility. Therefore $\mathbf{1}^\top \tilde{\mathbf{p}}_t = \mathbf{1}^\top \mathbf{d}^*_t$ for all t. Note that this is equivalent to $\sum_i \tilde{\mathbf{p}}_i = \sum_j \mathbf{d}^*_j$.

(ii) Total generation revenue is given by

$$\text{Rev}_{\text{gen}} = \sum_t \tilde{\pi}_t \mathbf{1}^\top \tilde{\mathbf{p}}_t = \sum_t \tilde{\pi}_t \mathbf{1}^\top \tilde{\mathbf{d}}_t$$

Total energy payments from demand are

$$\text{Pay}_{\text{demand}} = \sum_t \pi^0_t \mathbf{1}^\top \mathbf{d}^*_t = \sum_t \pi^0_t \mathbf{1}^\top \tilde{\mathbf{d}}_t$$

Total flexibility payments to loads are

$$\begin{aligned} \text{Rev}_{\text{flex}} &= \kappa^\top \sum_j \Delta^*_j \\ &= \sum_t \kappa_t \mathbf{1}^\top (\mathbf{d}^*_t - \mathbf{d}^0_t) \\ &= S \end{aligned}$$

Revenue neutrality is the condition when

$$\text{Pay}_{\text{demand}} - \text{Rev}_{\text{gen}} = \text{Rev}_{\text{flex}}.$$

By the definition of S, this condition is satisfied.

(iii) The proof of this result follows exactly the one for Theorem 3.2. The KKT stationarity condition for $\mathbf{p}_i$ is unaffected by the addition of constraints (9a) - (9c).

(iv) By primal feasibility of $\mathbf{d}^*_j$ in (11a - 11f), $d^*_{j,t} \geq d^0_{j,t}$ for $t \in \mathcal{T}$ and $d^*_{j,t} \leq d^0_{j,t}$ for $t \in \mathcal{T}^c$. In the proof of Lemma B.1 (see Appendix B) we showed that $\pi^0_t = \lambda^0_t = c_{\min}$ for $t \in \mathcal{T}$ and $c_{\min} < \pi^0_t$ for all $t \in \mathcal{T}^c$. As a consequence,

$$\sum_{t\in\mathcal{T}} \pi^0_t d^*_{j,t} - \sum_{t\in\mathcal{T}} \pi^0_t d^0_{j,t} = c_{\min} \sum_{t\in\mathcal{T}} (d^*_{j,t} - d^0_{j,t})$$

By primal feasibility of $\mathbf{d}^*_j$ and $\mathbf{d}^0_j$ from constraints (1c) and (11b) we have that

$$\begin{aligned} &\sum_t d^*_{j,t} = E_j = \sum_t d^0_{j,t} \\ \Rightarrow &\sum_{t\in\mathcal{T}} d^*_{j,t} + \sum_{t\in\mathcal{T}^c} d^*_{j,t} = \sum_{t\in\mathcal{T}} d^0_{j,t} + \sum_{t\in\mathcal{T}^c} d^0_{j,t} \\ \Rightarrow &\sum_{t\in\mathcal{T}} (d^*_{j,t} - d^0_{j,t}) = -\sum_{t\in\mathcal{T}^c} (d^*_{j,t} - d^0_{j,t}) \end{aligned}$$

Then

$$\begin{aligned} c_{\min} \sum_{t\in\mathcal{T}} (d^*_{j,t} - d^0_{j,t}) &= -c_{\min} \sum_{t\in\mathcal{T}^c} (d^*_{j,t} - d^0_{j,t}) \\ &\geq -\sum_{t\in\mathcal{T}^c} \pi^0_t (d^*_{j,t} - d^0_{j,t}) \end{aligned}$$

[10] A feasible solution exists: observe that $\tilde{\mathbf{D}}$ satisfies constraints (11b - 11f). An optimal value is attained because the objective function is continuous and the feasible set is compact.

By construction of κ, $\kappa_t \Delta_t^j \geq 0 \; \forall j, t$. Putting everything together,

$$\begin{aligned}
u_j^*(\mathbf{d}_j^*; \pi^0, \kappa) &= U_j - \sum_{t \in \mathcal{T}} c_{\min} d_{j,t}^* - \sum_{t \in \mathcal{T}^c} \pi_t^0 d_{j,t}^* + \sum_t \kappa_t \Delta_{j,t}^* \\
&\geq U_j - \sum_{t \in \mathcal{T}} c_{\min} d_{j,t}^* - \sum_{t \in \mathcal{T}^c} \pi_t^0 d_{j,t}^* \\
&= U_j - \sum_{t \in \mathcal{T}} c_{\min} d_{j,t}^* - \sum_{t \in \mathcal{T}^c} \pi_t^0 d_{j,t}^* \\
&\quad + \left(\sum_{t \in \mathcal{T}} c_{\min} d_{j,t}^0 + \sum_{t \in \mathcal{T}^c} \pi_t^0 d_{j,t}^0 \right) \\
&\quad - \left(\sum_{t \in \mathcal{T}} c_{\min} d_{j,t}^0 + \sum_{t \in \mathcal{T}^c} \pi_t^0 d_{j,t}^0 \right) \\
&= U_j - \sum_{t \in \mathcal{T}} c_{\min} (d_{j,t}^* - d_{j,t}^0) - \sum_{t \in \mathcal{T}^c} \pi_t^0 (d_{j,t}^* - d_{j,t}^0) \\
&\quad - \left(\sum_{t \in \mathcal{T}} c_{\min} d_{j,t}^0 + \sum_{t \in \mathcal{T}^c} \pi_t^0 d_{j,t}^0 \right) \\
&\geq U_j - \sum_{t \in \mathcal{T}} c_{\min} d_{j,t}^0 + \sum_{t \in \mathcal{T}^c} \pi_t^0 d_{j,t}^0 \\
&= U_j - \sum_t \pi_t^0 d_t^{j,0} \\
&= u_j(\mathbf{d}^0; \pi^0)
\end{aligned}$$

(v) Let $\mathcal{M}_{\text{flex}} \subseteq \{1, \ldots, M\}$ be the index set of loads whose flexibility is dispatched. Complementarily, $\mathcal{M}_{\text{flex}}^c \subseteq \{1, \ldots, M\}$ is the index set of loads whose flexibility is *not* dispatched. For $j \in \mathcal{M}_{\text{flex}}^c$,

$$\begin{aligned}
u_j^*(\mathbf{d}_j^*; \pi^0) &= U_j - {\pi^0}^\top \mathbf{d}_j^* \\
&\leq U_j - \tilde{\pi}^\top \tilde{\mathbf{d}}_j \\
&= u_j(\tilde{\mathbf{d}}_j; \tilde{\pi}).
\end{aligned} \tag{19}$$

The inequality comes from the fact proved in Lemma B.2.

Finally we show the second statement in (v):

$$\begin{aligned}
\sum_j u_j^*(\mathbf{d}_j^*; \pi^0, \kappa) &= \sum_{j \in \mathcal{M}_{\text{flex}}} u_j^*(\mathbf{d}_j^*; \pi^0, \kappa) + \sum_{j \in \mathcal{M}_{\text{flex}}^c} u_j^*(\mathbf{d}_j^*; \pi^0) \\
&= \sum_j U_j - \text{Pay}_{\text{demand}} + \text{Rev}_{\text{flex}} \\
&= \sum_j U_j - \text{Rev}_{\text{gen}} \\
&= \sum_j U_j - \sum_j \tilde{\pi}^\top \tilde{\mathbf{d}}_j \\
&= \sum_{j \in \mathcal{M}_{\text{flex}}} u_j(\tilde{\mathbf{d}}_j; \tilde{\pi}) + \sum_{j \in \mathcal{M}_{\text{flex}}^c} u_j(\tilde{\mathbf{d}}_j; \tilde{\pi})
\end{aligned}$$

From (19) we have that $\sum_{j \in \mathcal{M}_{\text{flex}}^c} u_j(\tilde{\mathbf{d}}_j; \tilde{\pi}) \geq \sum_{j \in \mathcal{M}_{\text{flex}}^c} u_j^*(\mathbf{d}_j^*; \pi^0, \kappa)$. The sequence of equalities implies

$$\sum_{j \in \mathcal{M}_{\text{flex}}} u_j^*(\mathbf{d}_j^*; \pi^0, \kappa) \geq \sum_{j \in \mathcal{M}_{\text{flex}}} u_j(\tilde{\mathbf{d}}_j; \tilde{\pi}).$$